NLP 2

Spracovanie prirodzeného jazyka

Peter Bednár

Textové dátové zdroje

Dátové zdroje (1)

- Text
  - dokumenty, články, knihy, e-maily, správy a príspevky na sociálnych sieťach
  - .txt, .md
- Formátované dokumenty
  - správy, zmluvy, prezentácie, elektronické knihy a súborové archívy
  - .docx, .pdf, .epub, .zip
- Štruktúrované a pološtruktúrované údaje
  - tabuľky, formuláre, databázy, webové stránky a služby
  - .json, .html, .xml
- Zdrojový kód
  - programy, skripty, komentáre a konfiguračné súbory

Dátové zdroje (2)

- Obraz
  - skenované dokumenty, fotografie, snímky obrazovky, diagramy a rukou písané poznámky
  - obraz → OCR → text
- Zvuk
  - reč, telefonické hovory, rozhovory, podcasty a hlasové správy
  - zvuk → rozpoznávanie reči → text
- Video
  - reč, titulky, text zobrazený v obraze a posunkový jazyk
  - zvuk/obraz → rozpoznávanie reči/obrazu → text

- Multimodálne údaje
  - kombinácie textu, obrazu, zvuku a videa v četoch, sociálnych sieťach a interaktívnych aplikáciách

Reprezentácie textu

Reprezentácie textu

Diskrétne	postupnosť znakov, postupnosť tokenov
Symbolické	závislostné stromy, znalostné grafy
Frekvenčné	*Bag of Words*, TF-IDF, BM25
Latentné	LSI, LDA
Distribuované	Word2Vec, GloVe, Doc2Vec
Kontextové	ElMo, BERT, GPT

Diskrétne reprezentácie textu

Reprezentácia textu na úrovni znakov

- Text je postupnosť znakov zakódovaných číselne

- UNICODE
  - viac než 297 000 znakov (písmen, číslic, interpunkcie, symbolov a emoji) pokrývajúcich 127 písem
  - UTF-8, 16, 32 – spôsob uloženia v bytoch

Písma (1)

- Abeceda
  - samostatné znaky reprezentujú spoluhlásky aj samohlásky
  - latinka, cyrilika, grécka abeceda
- Abdžad
  - zapisujú sa hlavne spoluhlásky, samohlásky sa často vynechávajú alebo označujú doplnkovými znamienkami
  - arabské písmo: پدر / پِدَر (⟨pedar⟩ - otec)

Písma (2)

- Abugida
  - jeden znak reprezentuje jednu spoluhlásku doplnenú samohláskou
  - dévanágarí (Hindi): क (ka), कि (ki), कु (ku), के (ke), को (ko)
- Slabičné písmo
  - jeden znak reprezentuje celú slabiku
  - japonská hiragana/katakana: さくら (sakura), カタカナ (katakana)
- Príznakové písmo
  - tvar znaku zodpovedá spôsobu tvorenia hlások
  - kórejský hangul: 카카오 (kakao)

Písma (3)

- Logografické písmo
  - jeden znak reprezentuje slovo alebo významovú jednotku
  - čínske znaky: 人 (človek), 山 (hora)
- Ideografické a symbolické systémy
  - znaky vyjadrujú pojmy nezávisle od konkrétneho jazyka
  - emotikony, číslice, matematické znaky, piktogramy
  - 😀 1 + 📵

Reprezentácia na úrovni tokenov

- Tokenizácia
  - prevedenie postupnosti znakov na postupnosť základných jazykových jednotiek – tokenov
- Rozdelenie na slová
  - regulárne výrazy, pravidlá na rozdeľovanie a spájanie slov, slovníky
  - modely strojového učenia
- Rozdelenie na podreťazce
  - je potrebný trénovací korpus (môže byť aj viacjazyčný)
  - z korpusu sa vyextrahujú často sa vyskytujúce podreťazce, jedno slovo môže byť rozdelené na viacero tokenov
  - BPE (*Byte Pair Encoding*), WordPiece
  - Unigram model

Tokenizácia na slová (1)

- Príklady regulárnych výrazov:

```
[a-záéíĺóŕúýčďľňšťžäô]+
```

  - slovo v slovenskej abecede, všetky písmená malé, napr. vŕba

```
[0-9]+(\.[0-9]*)? (€|EUR)
```

  - číselná hodnota v eurách, napr. 10.25 EUR

```
[a-zA-Z0-9_\-]+(\.[a-zA-Z0-9_\-]+)*@([a-zA-Z0-9_\-]+\.)+[a-z]{2,6}
```

  - emailová adresa, napr. peter.bednar@tuke.sk

Tokenizácia na slová (2)

- Výhody
  - tokeny čo najviac zodpovedajú gramatickým slovám
  - priamo interpretovateľné
- Nevýhody
  - jazykovo závislé, veľký rozmer slovníka, špeciálny token pre neznáme slová
  - problémy s viacjazyčnými textami a polysyntetickými jazykmi
    - elvégezhetitek
    - aannialiqpasaarama iksivaassuujaluarnikumut

Tokenizácia na podreťazce - BPE

- Vytvorí sa počiatočný slovník na úrovni znakov
- Iteratívne sa spájajú často sa vyskytujúce symboly do nových tokenov až kým sa nedosiahne stanovená veľkosť slovníka

- Príklad
- Pre slová 5 × cat 3 × cats 2 × car je počiatočný slovník: c a t s r
- Postupne spájame najfrekventovanejšie tokeny:

ca	10
at	8
ts	3
ar	2

slovník + ca

cat	8
ts	3
car	2

slovník + cat . . .

- Výsledný slovník: c a t s r ca cat ts

Tokenizácia na podreťazce - WordPiece

- Spája dvojice, ktoré sa spolu vyskytujú častejšie, než by sme očakávali
- Skóre odvodené z pravdepodobnosti:

$$score(a,b) \propto \frac{P(a,b)}{P(a)P(b)} = \frac{freq(a,b)}{freq(a)freq(b)}$$

kde $P(a,b)$ je pravdepodobnosť výskytu dvojice tokenov vedľa seba, a $P(a)$, $P(b)$ sú pravdepodobnosti jednotlivých tokenov

- Príklad
- Pre slová 5 × cat 3 × cats 2 × car

ca	$10/(10.100)$	$= 0.2$	slovník + ts ...
at	$8/(10.8)$	$= 0.1$	
ts	$3/(8.3)$	$= 0.125$	
ar	$2/(10.2)$	$= 0.1$	

Tokenizácia na podreťazce - Unigram model (1)

1. Model začína s veľmi veľkým slovníkom
2. Text sa rozdelí na najpravdepodobnejšiu postupnosť tokenov podľa aktuálneho slovníka podľa:

$$P(T) = \prod_{t_i \in T} t_i$$

3. Zachovajú sa iba najviac vyskytujúce sa tokeny

Tokenizácia na podreťazce - Unigram model (2)

- Príklad
- Počiatočný slovník:

cat	0,05
ca	0,1
ts	0,2
c	0,1
a	0,15
t	0,12
s	0,2

- Slovo cats sa dá rozdeliť viacerými spôsobmi:

cat s	$0{,}05 \times 0{,}2$	$= 0{,}01$
ca ts	$0{,}1 \times 0{,}2$	$= 0{,}02$
ca t s	$0{,}1 \times 0{,}12 \times 0{,}2$	$= 0{,}0024$
c a ts	$0{,}1 \times 0{,}15 \times 0{,}2$	$= 0{,}003$
c a t s	$0{,}1 \times 0{,}15 \times 0{,}12 \times 0{,}2$	$= 0{,}00036$

- Pre efektívne nájdenie najpravdepodobnejšej tokenizácie sa používa Viterbiho algoritmus

Tokenizácia na úrovni bytov

- Umožňuje reprezentovať ľubovoľný text v UNICODE vrátane medzier
- (viacjazyčný) text → UTF-8 kódovanie → postupnosť bytov → BPE/WordPiece/Unigram → slovník tokenov
  - 256 tokenov pre samostatné byty + tokeny pre podreťazce (dlhšie postupnosti bytov/UTF-8 znakov)
- Samostatné tokeny pre podreťazce na začiatku, alebo vo vnútri slov oddelených medzerou
  - Napr. The cat ##s are play ##ing

Tokenizácia pre jazykové modely

- Veľkosť slovníka pre jazykové modely:

~30k	~50k	~100k	~130–150k	~200–260k
BERT, T5	GPT-2	GPT-4	Llama 3 / Mistral / Qwen	GPT-4o / GPT-5 / Gemma 3

- Nárast hlavne kvôli viacjazyčnosti modelov
- Väčší slovník umožňuje rozdeliť text na menej tokenov (väčší kontext), ale vyžaduje si viac parametrov modelu na ich reprezentáciu (*embedding* vrstva)

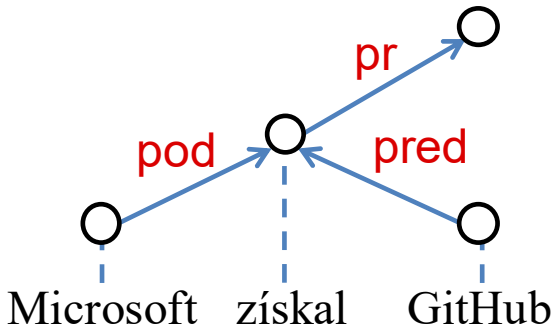
Symbolické reprezentácie textu

Morfosyntaktická reprezentácia

1. Tokenizácia textu na slová a segmentácia na vety
2. Morfologické značkovanie a lematizácia
   - priradí slovám základný tvar, slovný druh a morfologické kategórie
3. Syntaktické parsovanie
   - rozdelí vety na frázy, priradenie syntaktických rolí (podmet, prísudok, predmet, atď.)

Syntaktická reprezentácia

- Závislostné stromy
  - každému slovu vo vete zodpovedá jeden uzol + koreň stromu
  - orientované hrany medzi uzlami reprezentujú gramatické vzťahy a sú označené funkciou podradeného slova (počiatočný uzol) voči nadradenému slovu (koncový uzol hrany), napr.:

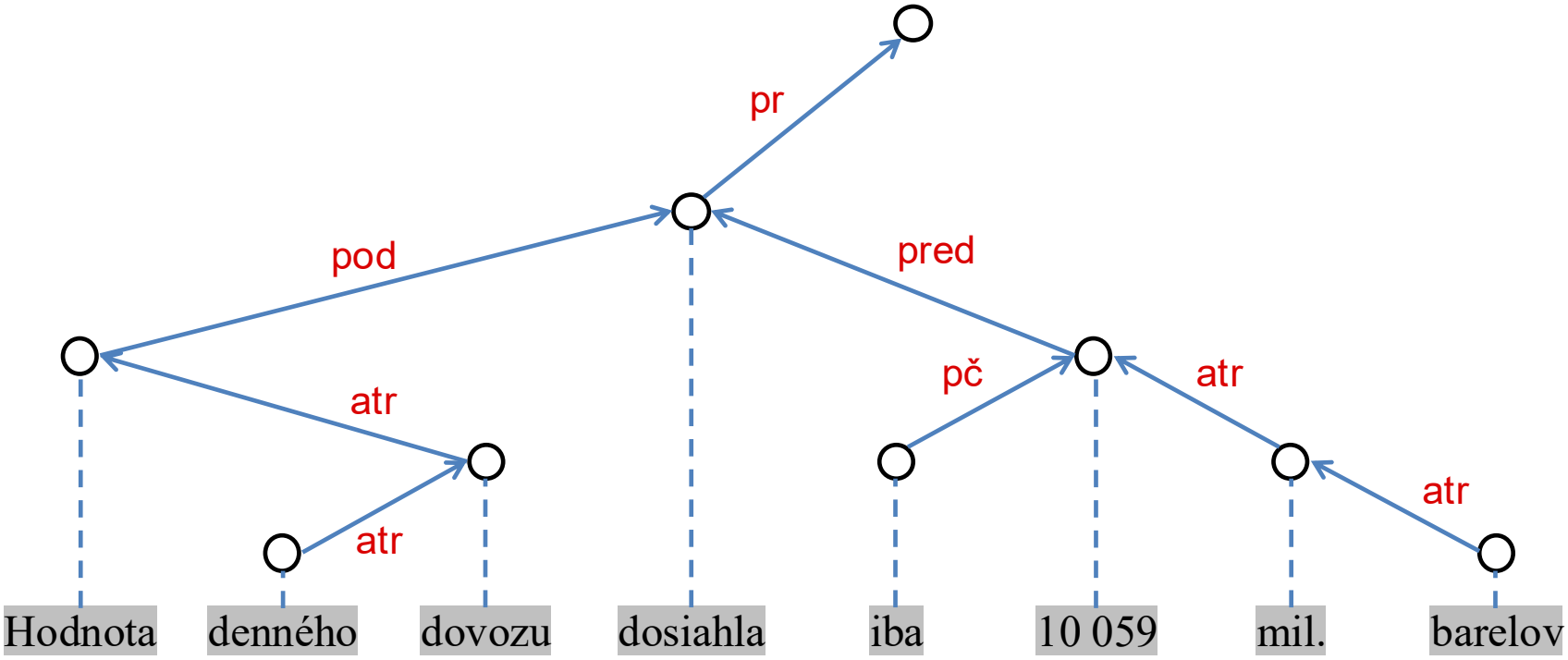


Morfosyntaktická reprezentácia – príklad (1)

- Slovné druhy
  - PM podstatné meno
  - PrM prídavné meno
  - S sloveso
  - ZČ zvýrazňovacia častica
  - Čsl číslovka
  - atď.
- Morfologické kategórie
  - 1p, 2p – pád
  - jč, mč – jednotné/množné číslo
  - žr, mr – ženský/mužský rod
  - nž – neživotné
  - dok – dokonavý slovesný vid
  - atď.

Hodnota	denného	dovozu	dosiahla	iba	10 059	mil.	barelov
PM 1p, jč, žr	PrM 2p, jč, mr	PM 2p, jč, mr, nž	S dok jč, žr	ZČ	Čsl	Čsl 2p, mč, mr	PM 2p, mč, mr, nž

Morfosyntaktická reprezentácia – príklad (2)



Syntaktické roly

pr – prísudok, pod – podmet, pred – predmet, atr – atribút rozvitého člena, pč – pomocný člen

Universal Dependencies

- Poskytuje zjednotenú definíciu morfologický kategórií a syntaktických rolí pre rôzne jazyky
  - 17 slovných druhov
  - 27 morfologických kategórií
- Anotované textové korpusy pre 150 jazykov
- Pre slovenčinu:
  - 106 K tokenov/slov, 10 604 anotovaných viet
- https://universaldependencies.org/

Sémantická reprezentácia

- Cieľom je vytvoriť jednoznačnú štruktúrovanú reprezentáciu významu textu
- Znalostných grafov
  - uzly – jednoznačne definované pojmy/entity
  - hrany – relácie medzi entitami
  - reprezentácia ako množina faktov v tvare subjekt-predikát-objekt – *tripletov*

Konštrukcia znalostných grafov

- Rozlíšenie významu slov (WSD)
- Rozpoznávanie pomenovaných entít (NER)
  - mestá, osoby, organizácie, udalosti ... produkty ... diagnózy, chemické zlúčeniny, DNA sekvencie, lieky ...
- Rozlíšenie koreferencií
  - ktoré výrazy v texte (názvy, skratky, zámená) označujú rovnakú identitu
- Extrakcia vzťahov medzi entitami

Karzai a Bush rokovali o Talibane. Americký prezident$_{[Bush]}$ sa stretol s afganským prezidentom$_{[Karzai]}$ na konferencii v Kábule.

⟨George H. W. Bush⟩ ⟨stretol sa⟩ ⟨Hamid Karzai⟩
...

Sémantická reprezentácia – príklad (1)

- Vety vyextrahované z medicínskych článkov

Stress is associated with migraines.
Stress can lead to loss of magnesium.
Calcium channel blockers prevent some migraines
Magnesium is a natural calcium channel blocker.
Spreading cortical depression (SCD) is implicated in some migraines.
High levels of magnesium inhibit SCD.
Migraine patients have high platelet aggregability.
Magnesium can suppress platelet aggregability.

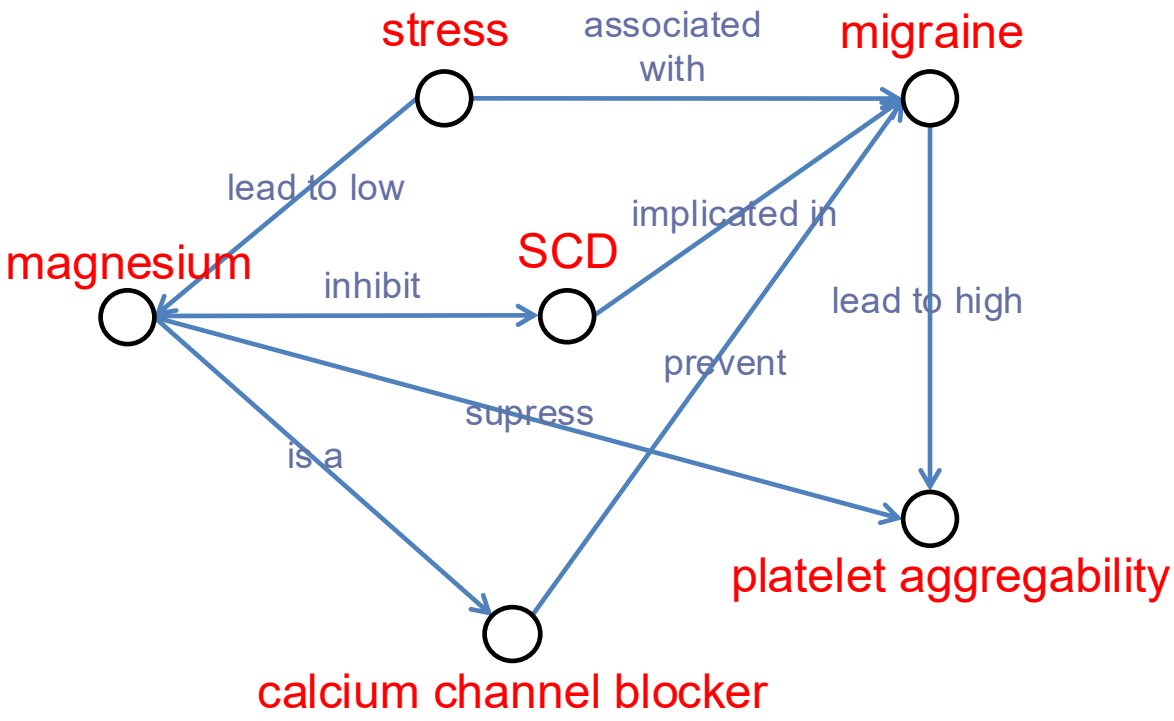
Sémantická reprezentácia – príklad (2)

- Extrahovanie entít

Stress is associated with migraines.
Stress can lead to loss of magnesium.
Calcium channel blockers prevent some migraines.
Magnesium is a natural calcium channel blocker.
Spreading cortical depression (SCD) is implicated in some migraines.
High levels of magnesium inhibit SCD.
Migraine patients have high platelet aggregability.
Magnesium can suppress platelet aggregability.

Sémantická reprezentácia – príklad (3)

- Extrahovanie relácií



Znalostné grafy – príklad (1)

- Wikidata
  - kolaboratívne vytváraný znalostný graf
  - relácie sú reprezentované tvrdeniami, ktoré okrem tripletu môžu obsahovať aj kvantifikátory a odkazy, napr.:

Barack Obama → funkcia → prezident USA
- začiatok → 20. 1. 2009
- koniec → 20. 1. 2017
- poradie → 44

  - viac ako 123 miliónov objektov, z toho ~13,7 milióna osôb, ~8,4 milióna astronomických objektov, ~6,1 milióna stavieb, ~3,9 milióna udalostí, ~441 tis. filmov, 1,2 mil. chemických zlúčenín, ~46 miliónov vedeckých článkov, ...
  - https://www.wikidata.org/

Znalostné grafy – príklad (2)

- Unified Medical Language System (UMLS)
  - systém integrujúci 195 biomedicínskych slovníkov a klasifikácií
  - 3,53 mil. hierarchicky usporiadaných pojmov
  - 127 typov entít, 54 relácií, terminológia v 31 jazykoch
  - ochorenia, diagnózy a symptómy; anatómia a fyziológia; lieky, účinné látky a liekové interakcie; laboratórne vyšetrenia a klinické merania; medicínske a chirurgické postupy; gény, proteíny a molekulárna biológia; mikroorganizmy a infekčné ochorenia; zdravotnícke pomôcky a materiály; ...
  - https://www.nlm.nih.gov/research/umls